COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI
DEPARTAMENTO DE MATEMÁTICA APLICADA

DMA-MBD-513 Fundamentals of Mathematical Data Analytics

SEMESTER: Fall

CREDITS: 30 hours

LANGUAGE: Spanish

DEGREES: Master in Big Data Technologies and Advanced Analytics

Course overview

This course is an introduction to statistics with a special emphasis on its mathematical foundations and a review of the main machine learning techniques. The subject aims to introduce the student to basic, but fundamental, concepts such as distributions, probability or inference. In addition, machine learning topics cover clustering, regression, forecasting and unsupervised learning. The subject also provides a first course of basic R by combining exposure of the main concepts in statistics and tutorial R sessions.

Prerequisites

Basic knowledge of Calculus and Algebra is required (understand and manipulate equations, manipulate exponents and logarithms using their basic rules, full understanding of functions and inverse functions, understand limits, derivatives and integrals, know rules for product and summation, etc.) It is also required basic knowledge of Statistics (descriptive statistics, discrete and continuous probability distribution models, sampling and basics of statistical inference).

Basic knowledge of Programming languages is required, ideally in R or Python.

Course contents

The contents of the course are divided in two sections: Statistics and Machine Learning

Statistics:

1. Introduction: Types of variables. Levels of measurement. Frequency tables.
2. Graphics: For categorical data. For quantitative data. Relationship between variables.
3. Distributions: Shape of distributions. Measures of centrality and dispersion. Summaries and boxplots.

4. Probability I: Contingency tables. The rules of probability. Conditional probability. Relationship between categorical variables.
5. Probability II: Normal. Properties of the normal distribution. Poisson. Weibull. Exponential. Sampling. Confidence Intervals. Central limit theorem.
6. Introduction to Bayesian Statistics: Bayes Theorem. Monte Carlo. Bayesian software.

Machine Learning

1. Classification: The classification problem. Cross-validation. kNN and decision trees.
2. Regression: The regression problem. Linear regression and neural networks.
3. Forecasting Time series decomposition. ARIMA models. Non-linear forecasting.
4. Unsupervised learning: PCA, hierarchical clustering, kmeans and density estimation.

Textbooks

While we will not follow a textbook, we find the following books quite remarkable in their central topics (R, regression and Bayesian statistics, respectively).

- **Kabacoff, R.,** (2011). *R in Action*. 1st Edition. Manning Publications.
- **Gelman, A. and Hill, J.,** (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models,* 1st Edition. Cambridge University Press.
- **Kruschke, J.,** (2014), *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan,* 2nd Edition. Academic Press.
- **G. James, D. Witten, T. Hastie & R. Tibshirani** (2013). An Introduction to Statistical Learning with Applications in R. Springer (see http://www-bcf.usc.edu/~gareth/ISL/)
- **W. Wei** (2006). *Time Series Analysis. Univariate and Multivariate Methods.* 2nd Ed.

The following conditions must be accomplished to pass the course:

- A minimum overall grade of at least 5 over 10.
- A minimum grade in the final exam of 4 over 10.

The overall grade is obtained as follows:

- Final exam accounts for 60% of the final grade if the grade in this exam is at least 4. In other case, final exam accounts for 100 % of the overall grade.
- Laboratory sessions work (in class and homework) accounts for 40% of the final grade.